The Meaning of Word Sense Disambiguation Research

Postma, M.C.

2019

document version
Publisher's PDF, also known as Version of record

Chapter 16

Samenvatting

Woorden hebben veel betekenissen en betekenissen kunnen uitgedrukt worden door veel woorden. Dat is wat bestudeerd wordt in dit proefschrift. Mensen zijn heel goed in het interpreteren van taal, terwijl computers hier meer moeite mee hebben.

In een poging om computers beter te laten worden in het interpreteren van taal hebben onderzoekers ervoor gekozen om het complexe probleem op te knippen in stukken. Een voorbeeld hiervan is de Word Sense Disambiguation taak, waarvan het doel is om aan te geven wat een woord betekent in een bepaalde zin, b.v. wat betekent het woord *paard* in de zin *met die zet van het paard, staat de tegenstander schaakmat*. Maar er zijn nog veel meer taken bedacht. Denk bijvoorbeeld aan het interpreteren van eigennamen, tijdsuitdrukkingen of het aangeven welke uitdrukkingen over dezelfde persoon of gebeurtenis gaan, b.v. dat *Jan* en *hij* naar dezelfde persoon verwijzen in de zin *Jan gaat naar huis en hij gaat een koekje eten*.

In dit proefschrift hebben we de Word Sense Disambiguation taak onderzocht. De taak is één van de oudste in het veld en vele aanpakken zijn er geprobeerd om de taak op te lossen, maar zijn hier niet in geslaagd. Wij vragen ons af waarom dit het geval is en wat we hiervan kunnen leren.

In deel 1 van de dissertatie introduceren we eerst de belangrijkste componenten van de Word Sense Disambiguation taak. Een digitaal woordenboek genaamd WordNet staat centraal in de taak aangezien dit aangeeft wat de betekenissen van woorden zijn alswel wat de relaties tussen betekenissen zijn. Gebruikmakend van WordNet zijn verder grote hoeveelheiden tekst "geannoteerd" door mensen, dat willen zeggen dat voor veel zinnen is aangegeven wat de woorden erin betekenen. Er zijn in de loop der jaren veel systemen gemaakt die proberen om automatisch de betekenis van een woord in een zin te voorspellen, gebruikmakend van zowel de informatie in WordNet alsook van de geannoteerde data. Een uitgebreide analyse van deze systemen leert ons dat in de loop der jaren de systemen niet veel beter zijn geworden. Daarbij zijn systemen veel beter in het herkennen van betekenissen die veel voorkomen dan in het herkennen van betekenissen die niet vaak voorkomen. Daarbij zit al ingebouwd in veel systemen dat ze de voorkeur geven aan betekenissen die veel voorkomen. Daarnaast hebben we kritisch gekeken naar de data die we gebruiken om aan te geven hoe goed een systeem is in het interpreteren van taal. We observeren dat deze data niet representatief is voor het normale taalgebruik, maar dat ze de veelvoorkomende woorden en betekenissen overrepresenteren.

In deel 2 kijken we kritisch naar de rol van geannoteerde data. Veel taalsystemen maken gebruiken van geannoteerde data om voorspellingen te doen over wat een woord betekent in een zin. Patronen worden herkend in de geannoteerde data en deze worden dan gebruikt op nieuwe zinnen om voorspellingen te doen over wat de woorden in die nieuwe zinnen betekenen. We manipuleren deze geannoteerde data om te kijken hoe deze taalsystemen reageren. We bekijken bijvoorbeeld de rol van de hoeveelheid geannoteerde data, de proportie niet populaire betekenissen, of dat de data door mensen is geannoteerd of door taalsystemen. We observeren dat meer data niet per se tot betere systemen leidt. Daarbij veranderen systemen drastisch

in hoe ze betekenissen toekennen als er relatief gezien meer onpopulaire betekenissen in de geannoteerde data zitten.

In deel 3 denken we na over een andere manier om te bepalen hoe goed taalsystemen zijn in het interpreteren van taal. De standaardmanier is om mensen data te laten annoteren. Een mens bepaalt bijvoorbeeld dat *paard* in de zin *het paard staat in de wei* de betekenis van het dier heeft in plaats van bijvoorbeeld de schaakbetekenis. Wij draaien de procedure om. We beginnen met een database van gebeurtenissen van wapengeweld, bijvoorbeeld incidenten waarbij mensen vermoord zijn. Veel gebeurtenissen hebben meerdere verschillende nieuwsberichten die deze gebeurtenissen beschrijven. Vervolgens voegen we al deze documenten samen die al deze gebeurtenissen beschrijven. De taak is dat we vragen hebben bedacht die het taalsysteem moet beantwoorden. Bijvoorbeeld, welke documenten beschrijven gebeurtenissen van moorden die in 2013 plaatsgevonden hebben in de staat New York? De taak van het systeem is om de goede documenten te vinden. Om dit accuraat te doen, moet het systeem rekening houden met een groot aantal aspecten, waarbij het niet kan terugvallen op het voortrekken van dominante fenomenen.